Project report on

Malignant Comment Classifier

Submitted by:

Sri Ram Prasad Commuri

ACKNOWLEDGMENT

I would like to express my sincere thanks of gratitude to my SME Gulshana Chaudhary and "Flip Robo Technologies" team for providing an opportunity and allowing me to work on " Malignant comment classifier " project, also, thanks to academic team "Data Trained". Constant suggestions and directions from both SME , Flip Robo have helped me in completion of this project . This project helped me with lots of research knowledge wherein I explored many new things. All the required information and dataset are provided by Flip Robo Technologies (Bengaluru) for successful project completion.

The following research papers helped me understand the topic of malignant comments, what are the different parameters which help define a comment as malignant, the process of building a malignant comments classifier & finally, helped me in my model building & predictions:

1. "Toxic Comment Classification"- Sara Zaheri, Jeff Leath, David Stroud

This paper presents a novel application of Natural Language Processing techniques to classify unstructured text into toxic and nontoxic categories. In the current century, social media has created many job opportunities and, at the same time, it has become a unique place for people to freely express their opinions. Meanwhile, among these users, there are some groups that are taking advantage of this framework and misuse this freedom to implement their toxic mindset (i.e insulting, verbal sexual harassment, threads, Obscene, etc.). The 2017 Youth Risk Behavior Surveillance System (Centers for Disease Control and Prevention) estimated that 14.9% of high school students were electronically bullied in the 12 months, prior to the survey. The primary result could be an Open Source model used by app developers in support of anti bullying efforts.

2. "Identification and Classification of Toxic Comment Using Machine Learning Methods" - P.Vidyullatha, Satya Narayan Padhya, Javvaji Geetha Priya, Kakarlapudi Srija, Sri Satyanjani Koppisetti

The increase in penetration of usage of internet services has increased exponentially in the past 4 months due to the ongoing pandemic, this has empowered an enormous number of dynamic new and old clients utilizing the web for different administration ranging from academic, entertainment, industrial, monitoring and the emergence of a new trend in the corporate life i.e work-from-home. Due to this sudden emergence of the crowd using the web, there has been an ascent in the number of mischievous persons too. Now it is the primary task of every online platform provider to keep the conversations constructive and inclusive. The best example can be referred to, can be twitter, a web-based media stage where people share their views. This platform has already drawn a lot of flak because of the spread of hate speech, insults, threats, and defamatory acts which becomes a challenge for many such online providers in regulating them. Thus, there is active research being conducted in the field of Toxic comment classification.

3. "A Machine Learning Approach to Comment Toxicity Classification" - Navoneel Chakraborty

Now-a-days, derogatory comments are often made by one another, not only in offline environments but also immensely in online environments like social networking websites and online communities. So, an Identification combined with Prevention System in all social networking websites and applications, including all the communities, existing in the digital world is a necessity. In such a system, the Identification Block should identify any negative online behavior and should signal the Prevention Block to take action accordingly. This study aims to analyze any piece of text and detect different types of toxicity like obscenity, threats, insults and identity-based hatred. The labeled Wikipedia Comment Dataset prepared by Jigsaw is used for the purpose.

4. "An Automated Toxicity Classification on Social Media Using LSTM and Word Embedding" - Ahmad Alsharef, Karan Aggarwal, Sonia, Deepika Koundal, Hashem Alyami, and Darine Ameyed

The automated identification of toxicity in texts is a crucial area in text analysis since the social media world is replete with unfiltered content that ranges from mildly abusive to downright hateful. Researchers have found an unintended bias and unfairness caused by training datasets, which caused an inaccurate classification of toxic words in context. In this paper, several approaches for locating toxicity in texts are assessed and presented aiming to enhance the overall quality of text classification. General unsupervised methods were used depending on the state-of-art models and external embeddings to improve the accuracy while relieving bias and enhancing F1-score. Suggested approaches used a combination of long short-term memory (LSTM) deep learning model with Glove word embeddings and LSTM with word embeddings generated by the Bidirectional Encoder Representations from Transformers (BERT), respectively.

INTRODUCTION

Business Problem Framing

The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection. Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behavior.

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts.

Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it.

The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as unoffensive, but “u are an idiot” is clearly offensive.

Conceptual Background of the Domain Problem

In the past few years, it is seen that the cases related to social media hatred have increased exponentially. Social media is turning into a dark venomous pit for people nowadays. Online hate is the result of differences in opinion, race, religion, occupation, nationality etc. In social media the people spreading or involved in such kinds of activities use filthy language, aggression, images etc. to offend and gravely hurt the person on the other side. This is one of the major concerns now.

The result of such activities can be dangerous. It gives mental trauma to the victims making their lives miserable. People who are not well aware of mental health online hate or cyberbullying become life threatening for them. Such cases are also at rise. It is also taking its toll on religions. Each and every day we can see an incident of fighting between people of different communities or religions due to offensive social media posts.

Online hate, described as abusive language, aggression, cyberbullying, hatefulness, insults, personal attacks, provocation, racism, sexism, threats, or toxicity has been identified as a major threat on online social media platforms. These kinds of activities must be checked for a better future.

Motivation for the Problem Undertaken

The main objective of this study is to investigate which method from a chosen set of machine learning techniques performs the best. So far, we have a range of publicly available models served through the Perspective API, including toxicity/malignant comments. But the current models still make errors, and they don’t allow users to select which type of toxicity they are interested in finding.

The project gives an insight to identify major factors that lead to cyberbullying and online abusive comments. The exposure to real world data and the opportunity to deploy my skills in solving a real time problem has been the primary objective. However, the motivation for taking this project was that it is relatively a new field of research.

The main motivation was to classify the news in order to bring awareness and reduce unwanted chaos and make a good model which will help us to know such kinds of miscreants. Our goal is to build a prototype of online hate and abuse comment classifier which can be used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

ANALYTICAL PROBLEM FRAMING

Mathematical/Analytical Modeling of the Problem

We are provided with two different datasets. One for training and another one to test the efficiency of the model created using the training dataset. The training data provided here has both dependent and independent variables. As it is a multiclass problem it has 6 independent/target variables. Here the target variables are named “malignant”, “highly malignant”, “rude”, “threat”, “abuse” and “loathe”. The label can be either 0 or 1, where 0 denotes a NO while 1 denotes a YES. There are various comments which have multiple labels. The first attribute is a unique ID associated with each comment.

Clearly it is a binary classification problem as the target columns giving binary outputs and all independent variables have text so it is clear that it is a supervised machine learning problem where we can use the techniques of NLP and classification-based algorithms of Machine Learning. Here we will use NLP techniques like word tokenization, lemmatization, stemming and tfidf vectorizer then those processed data will be used to create best model using various classification based supervised ML algorithms like Logistic Regression, Multinomial NB, Extra Trees Classifier, XGBoost Classifier, LinearSVC, Decision Tree Classifier and Adaboost Classifier.

Data Sources & their formats

Data set provided by Flip Robo was in the format of CSV (Comma Separated Values). The data set contains the training set, which has approximately 159571 samples and the test set which contains nearly 153164 samples.

FEATURE INFORMATION:

1. Malignant: It is the Label column, which includes values 0 and 1, denoting if the comment is malignant or not.
2. Highly Malignant: It denotes comments that are highly malignant and hurtful.
3. Rude: It denotes comments that are very rude and offensive.
4. Threat: It contains indications of the comments that are giving any threat to someone.
5. Abuse: It is for comments that are abusive in nature.
6. Loathe: It describes the comments which are hateful and loathing in nature.
7. ID: It includes unique Ids associated with each comment text given.
8. Comment text: This column contains the comments extracted from various social media platforms.

Data preprocessing done

- Importing necessary libraries and loading the train and test datasets.
- Checked some statistical information like shape, number of unique values present, info, null values, value counts, duplicated values etc.
- Checked for null values and did not find any null values. And removed Id.
- Done feature engineering and created new columns viz label: which contain both good and bad comments which is the sum of all the labels, comment_length: which contains the length of comment text.
- Visualized each feature using seaborn and matplotlib libraries by plotting categorical plots like pie plot, count plot, distribution plot and word cloud for each label.
- Done text pre-processing techniques like Removing Punctuations and other special characters, Splitting the comments into individual words, Removing Stop Words, Stemming and Lemmatization.
- Then created a new column as clean_length after cleaning the data. All these steps were done on both train and test datasets.
- Checked correlation using heatmap.
- After getting cleaned data use TF-IDF vectorizer. It’ll help to transform the text data to feature vectors which can be used as input in our modeling. It is a common algorithm to transform text into numbers. It measures the originality of a word by comparing the frequency of appearance of a word in a document with the number of documents the words appear in.Mathematically, TF-IDF = TF(t*d)*IDF(t,d)
- Balanced the data using Randomoversampler method.
- Proceeded with Model Building.

Data Inputs- Logic- Output Relationships

The dataset consists of a label and features. The features are independent and the label is dependent as the values of our independent variables change as our label varies.

● Checked the distribution of skewness using dist plots and used count plots to check the counts available in each column as a part of univariate analysis.
● To analyze the relation between features and label I have used many plotting techniques where I found some of the columns having strong relation with label.
● Got to know the sense of loud words in every label using word cloud which gives the words frequented in the labels.
● I have checked the correlation between the label and features using a heat map.

Hardware and Software Requirements & Tools used

To build the machine learning projects it is important to have the following hardware and software requirements and tools.
Hardware required:
● Processor: core i5 or above
● RAM: 8 GB or above
● ROM/SSD: 250 GB or above
Software required:
● Anaconda - language used Python 3
Libraries Used:
● import pandas as pd
● import numpy as np
● import seaborn as sns
● import matplotlib.pyplot as plt
● import warnings warnings.filterwarnings('ignore')
● import nltk
● import string
● from nltk.corpus import stopwords
● from wordcloud import WordCloud
● from nltk.tokenize import word_tokenize
● from nltk.stem import WordNetLemmatizer
● from sklearn.feature_extraction.text import TfidfVectorizer
● from wordcloud import WordCloud
● from sklearn.model_selection import train_test_split
● from imblearn.over_sampling import RandomOverSampler
● from collections import Counter
● from sklearn import metrics
● from sklearn.model_selection import cross_val_score
● from sklearn.model_selection import train_test_split
● from sklearn.metrics import classification_report, confusion_matrix
● from sklearn.metrics importroc_curve, accuracy_score, roc_auc_score, hamming_loss, log_loss.
● from sklearn.linear_model import LogisticRegression
● from sklearn.tree import DecisionTreeClassifier
● from sklearn.svm import LinearSVC
● from sklearn.naive_bayes import MultinomialNB
● from sklearn.ensemble import AdaBoostClassifier
● from xgboost import XGBClassifier
● from sklearn.ensemble import ExtraTreesClassifier
● from sklearn.model_selection import cross_val_score
● from sklearn import datasets
● from sklearn import model_selection
● from sklearn.metrics import plot_roc_curve
● from sklearn.model_selection import GridSearchCV
● import joblib

MODEL BUILDING & EVALUATION

Identification of possible problem-solving approaches (methods)

I have used both statistical and analytical approaches to solve the problem which mainly includes the pre-processing of the data and also used EDA techniques and heat map to check the correlation of independent and dependent features. Also, before building the model, I made sure that the input data was cleaned and scaled before it was fed into the machine learning models.

In this project there were 6 features which defines the type of comment like malignant, hate, abuse, threat, loathe but we created another feature named as "label" which is combined of all the above features and contains the labeled data into the format of 0 and 1 where 0 represents "NO" and 1 represents "Yes".

I have used many classification algorithms and got the prediction results. By doing various evaluations I have selected the Logistic Regression Model as the best suitable algorithm to create our final model.

Testing of Identified Approaches (Algorithms)

Since the target variable is categorical in nature, from this I can conclude that it is a classification type problem hence I have used the following classification algorithms. After the pre-processing and data cleaning I left with 10 columns including targets.

The algorithms used on training the data are as follows:

1. Logistic Regression Model
2. Decision Tree Classifier Model
3. Linear SVC Model
4. MultinomialNB Classifier Model
5. AdaBoost Classifier Model
6. Extreme Gradient Boosting Classifier (XGB) Model
7. Extra Trees Classifier Model

Interpretation of the Results

Visualizations: I have used a distribution plot to visualize how the data has been distributed. Used count plots and pie charts to check the count of a particular category for each feature. The heat map helped me to understand the correlation between dependent and independent features. Also, heat maps helped to detect the multicollinearity problem and feature importance. With the help of Word Clouds I would be able to sense the loud words in each label. The AUC-ROC curve helped to select the best model.

Pre-processing: The dataset should be cleaned and scaled to build the ML models to get good predictions. I have performed a few NLP text processing steps which I have already mentioned in the pre-processing steps where all the important features are present in the dataset and ready for model building.

Model building: After cleaning and processing data, I performed a train test split to build the model. I have built multiple classification models to get the accurate accuracy score, and evaluation metrics like precision, recall, confusion matrix, f1 score, log loss, hamming loss. I got the Logistic Regression Model as the best model which gives a 94.52% accuracy score. I checked the cross-validation score ensuring there will be no overfitting. After tuning the best model, I got a 94.56% accuracy score. Finally, I saved my final model and got the good predictions results for the test dataset.

CONCLUSION

Key Findings and Conclusions of the Study

● From the above analysis the below mentioned results were achieved which depicts the chances and conditions of a comment being a hateful comment or a normal comment
● With the increasing popularity of social media more and more people consume feeds from social media and due differences they spread hate comments instead of love and harmony. It has strong negative impacts on individual users and broader society.
● From this dataset we were able to understand the impact of various malignant comments, the different types of malignant comments and how to identify them.
● In this study, we have used multiple machine learning models to predict malignant comments and identify them.
● We have gone through the data analysis by performing feature engineering, finding the relation between features and the label through visualizations. And got the important features and we used these features to predict the defaulters' rate by building ML models.
● After training the model we checked CV score to overcome the overfitting issue.
● Performed hyper parameter tuning, on the best model. We have also got good prediction results.

Learning Outcomes of the Study in respect of Data Science

While working on this project I learned many things and gained new techniques and ways to deal with uncleaned text data. Found how to deal with multiple target features. Tools used for visualizations give a better understanding of the dataset. We have used a lot of algorithms and find that in the classification problem where we have only two labels, the Logistic Regression Model gives better results compared to others.

Limitations of this work and Scope for Future Work

LIMITATIONS:
● This project was amazing to work on, it creates new ideas to think about but there were some limitations in this project like unbalanced dataset. There are certain areas which can be enhanced.

FUTURE WORK:
● The potential future work for this project can focus on performance and error analysis of the model as lots of comments are misclassified into the hate category.
● Previous work has achieved success using various algorithms on data in the English language but in future, we can consider having data in regional languages.